TITLE: Responsiveness and Minimal Important Change for Pain and Disability Outcome Measures in Pregnancy-Related Low Back and Pelvic Girdle Pain

RUNNING HEAD: Responsiveness of Pregnancy-Related LBP Measures

ARTICLE TYPE: Original Research

SECTION/TOC CATEGORY: Women's Health

AUTHOR BYLINE: Reuben Ogollah, Annette Bishop, Martyn Lewis, Margreth Grotle, Nadine E. Foster

AUTHOR INFORMATION:
R. Ogollah, PhD, Nottingham Clinical Trials Unit, School of Medicine, University of Nottingham, D Floor, South Block, Queens Medical Centre, Nottingham NG7 2UH, United Kingdom. Address all correspondence to Dr Ogollah at: reuben.ogollah@nottingham.ac.uk.

A. Bishop, PT, PhD, Research Institute for Primary Care and Health Sciences, Keele University, Keele, Staffordshire, United Kingdom.

M. Lewis, PhD, Research Institute for Primary Care and Health Sciences, Keele University; and Keele Clinical Trials Unit, Keele University.

M. Grotle, PT, PhD, Department of Physiotherapy, Oslo Metropolitan University, Oslo, Norway; and FORMI, Oslo University Hospital, Oslo, Norway.

N.E. Foster, PT, BSc(Hons), FCSP, DPhil, Research Institute for Primary Care and Health Sciences, Keele University; and Keele Clinical Trials Unit, Keele University.

KEYWORDS: Pregnancy-Related Low Back Pain and Pelvic Girdle Pain, Minimally Important Change, Responsiveness, Patient-Based Outcome Measures

ACCEPTED: March 25, 2019

SUBMITTED: July 10, 2018

Background. Pregnancy-related low back pain and pelvic girdle pain (LBP/PGP**)** are common and negatively impact the lives of many pregnant women. Several patient-based outcome instruments measure treatment effect but there is no consensus about which measure to use with women who have these pain presentations.

Objective. The objective was to compare the responsiveness of 3 outcome measures in LBP/PGP: Oswestry Disability Index-version 2.0 (ODI), Pelvic Girdle Questionnaire (PGQ), and 0-10 numerical rating scale for pain severity (NRS); and to estimate a minimal important change (MIC) for these measures in pregnancy-related LBP/PGP.

Design. This was a methodology study using data from a pilot randomised trial (RCT).

Methods. Women (n = 124) with pregnancy-related LBP/PGP were recruited to a pilot RCT evaluating the benefit of adding acupuncture to standard care and 90 completed 8-weeks follow-up. Responsiveness was evaluated by examining correlation between change score and the external anchor (6-point global perceived effect scale) and by using receiver operating characteristic (ROC) curve analysis. MIC was estimated using anchor-based methods.

Results. All measures showed good responsiveness, with areas under ROC curve ranging from 0.77 to 0.90. The estimated MICs were 3.1, 11.0, 9.4, 13.3, and 1.3 for ODI, PGQ-total, PGQ-activity, PGQ-symptoms, and NRS, respectively. All the measures, apart from ODI, had MICs larger than the measurement error.

Limitations. Lack of optimal "gold standard" or external criterion for assessing responsiveness and MIC was a limitation of this study.

Conclusion. All 3 outcome measures demonstrated good responsiveness. MICs were derived for each instrument. The PGQ at 8 weeks post-randomisation was identified as an appropriate outcome measure for pregnancy-related LBP/PGP since it is specific to these pain presentations and assesses both activity limitations and symptoms. The NRS is an efficient, shorter alternative.

Pregnancy-related low back pain (LBP) and pelvic girdle pain (PGP) are common conditions[1,2] and impact, both socially and economically, the daily lives of many pregnant women.[3] The incidence and point prevalence of pregnancy-related LBP and PGP reported in literature vary, partly because there is a variation in the criteria employed by various studies for the diagnosis of these conditions and the designs of the studies, but estimates range from 24% to 90% of women experiencing LBP or PGP at some stage during their pregnancy.[1,4-6] The two pain presentations may occur separately or together and typically increase with advancing pregnancy.[2]

Treatment of pregnancy-related LBP and PGP is often aimed at reducing the pain and improving function. Therefore, assessing whether pain or function has changed over time is a vital objective of measurements in clinical practice and health research. Outcome measures for LBP and for the disability caused by LBP have become important standards for evaluation of interventions due to their frequent use in the assessment of the effectiveness of treatments.[7] There have been several consensus developing initiatives to agree on core outcome domains and core outcome measures for non-specific LBP,[7,8] however, there is a lack of consensus on what measure to use for pregnancy-related LBP and PGP as outcome measures that have been validated for non-specific LBP are not necessarily the most appropriate for pregnancy-related LBP and PGP.[9] There is therefore a need for a direct comparison between LBP outcome measures in order to inform the choice of the most appropriate outcome measure to use in pregnancy-related LBP and PGP.

An important measurement property of a patient reported outcome measure (PROM) is its ability to detect clinical change over time. This can be evaluated by determining the responsiveness of the PROM, defined as the ability of an instrument to detect change over time in the construct to be measured.[10] A related question of interest is whether the observed changes within patients are clinically relevant. This can be evaluated by determining the minimal important change (MIC), defined as the minimal amount of change that is important to the patient.[11,12] Analysing the responsiveness and MICs of an outcome measure is a continuous process that is strongly recommended to strengthen its properties and expand its applicability.[13]

Since two main aims of treatment for pregnancy-related LBP and PGP are to relieve pain and to improve functional ability, we included PROMs that are frequently used to capture pain and impact of pain on everyday activities (measured using both the Oswestry Disability Index and the Pelvic Girdle Questionnaire (PGQ). The responsiveness for the ODI and NRS has been established for the non-specific LBP population, but not in pregnancy-related LBP or PGP. The Pelvic Girdle Questionnaire (PGQ) is a recent outcome measure developed for both pregnant and postpartum women and its responsiveness and MIC have been established for women in late pregnancy and transitioning to postpartum,[14] however, no previous study has established the properties of this instrument entirely during the course of pregnancy within a trial context. The previous study[14] recruited women in the last trimester of their pregnancy and within 3 months after delivery and therefore their assessment of the measurement properties are more relevant to the context of natural history or epidemiological cohort studies rather than clinical trials **of** treatment for LBP and PGP during pregnancy.

In order to aid decision-making about outcome measures for use in clinical trials of treatment effectiveness for pregnancy-related LBP and PGP, it is important to determine the responsiveness and MICs of the PROMs used to capture pain and impact of pain in women seeking treatment. This can be best done by choosing a population of women seeking treatment for pregnancy-related LBP and PGP and a time-point during pregnancy in the evaluation of responsiveness and MICs.

The purpose of this study was twofold: (1) to compare the responsiveness of three PROMs during pregnancy in patients with pregnancy-related LBP and PGP: the Oswestry Disability Index-version 2.0 (ODI), the Pelvic Girdle Questionnaire (PGQ), and a 0-10 numerical rating scale (NRS) for pain severity, and (2) to estimate the optimal cut-off values for a MIC for these measures with patients who have pregnancy-related LBP and PGP.

[H1]Methods

[H2]Study participants

This study used the data from a three-arm parallel pilot randomised controlled trial (RCT) that assessed the feasibility of a future large RCT testing the additional benefit of acupuncture to standard

care for pregnancy-related LBP and PGP– the EASE Back pilot trial[15] (trial registration ISRCTN49955124). Full details of the EASE Back pilot trial are available elsewhere,[15,16] including the full details of the definitions of LBP and PGP we used. Here we give brief details. The EASE Back trial was funded following a commissioned call by the UK's National Institute for Health Research (NIHR) Health Technology Assessment programme, focused on pregnancy-related LBP with or without PGP. Women aged 18 years and over were included if they had pregnancy-related LBP defined as self-reported pain in the lumbar area (between the 12th rib and the gluteal fold) with or without PGP (defined by pain between the posterior iliac crest and the gluteal fold, particularly in the vicinity of the sacro-iliac joints), were under the care of participating NHS sites and GP practices, at 13 to 31 weeks gestation, were naïve to acupuncture treatment, able to read and communicate in English and were willing to participate. Women who had LBP episodes before this pregnancy were included as long as the current episode of LBP was either attributed to, or made worse as a result of, this pregnancy. Women were excluded if they had previously had any form of acupuncture, were at high risk of miscarriage or pre-term labour, had pre-eclampsia, previous history of surgery to the spine or pelvis, and had contra-indications to the treatments, pain in the anterior pelvic region only, or a current urinary tract infection. Eligible women who gave written informed consent to participate were randomised in a 1:1:1 ratio, to one of the three treatment arms: standard care alone (comprehensive self-management booklet and an option to access EASE Back physiotherapy care if needed); standard care plus a course of true acupuncture; and standard care plus a course of non-penetrating acupuncture, delivered by physiotherapists. Ethical approval for the EASE Back pilot trial was granted by NRES Committee West Midlands - Staffordshire (reference 13/WM/00).

[H2]Sample size

In line with pilot trial recommendations[17] no formal sample size calculation was conducted for the EASE Back pilot trial for the key clinical outcomes. However, we aimed to recruit a large number of patients to obtain an acceptable number of patients for the responsiveness and MIC analysis based on the guidelines for sample sizes in studies on validity and responsiveness.[13] A total of 124 pregnant

women were randomised and provided baseline data, while 90 also provided data on the three outcome measures at 8 weeks follow-up.

[H2]Data collection

At baseline, patients completed a questionnaire containing sociodemographic data and PROMs. At 8 weeks post-randomisation, patients completed the same outcome measures used at baseline assessment and a patient's Global Perceived Effect (GPE) scale for change of the pregnancy-related LBP/PGP and disability measured using a 6-point ordinal scale: (1) completely recovered, (2) much improved, (3) somewhat improved, (4) same, (5) somewhat worse, and (6) much worse. The decision to collect outcome data at 8 weeks post-randomisation was timed to ensure that most women would not yet have given birth to their baby.

[H2]Outcome measures

The EASE Back pilot trial did not have a primary clinical outcome. Instead, two outcome measures capturing data on physical function and one instrument of pain severity were included in the pilot trial, in order to help inform the decision about a primary outcome measure for a future main trial. Several factors were taken into consideration in examining the performance of these measures. These included the amount of missing data at the item and scale levels; precision of the outcome measures; any evidence of floor or ceiling effects; their responsiveness to change; and MICs. Evaluation of these factors is the focus of this paper. The PROMs used included:

[H3]Oswestry disability index (ODI), version 2.0

The ODI[18,19] is a self-administered questionnaire containing 10 sections, each containing six statements that are scored from 0 to 5, with 0 representing no difficulty in the activity and 5 representing maximal difficulty. The scores from each section are totalled and divided by the total possible score to obtain a final percentage of disability, with a higher percent indicating greater disability. Decreasing scores for the ODI over time denote improvement. The ODI is a valid, reliable, and responsive clinical tool for analysing disability status in individuals with LBP.[20]

[H3]Pelvic Girdle Questionnaire (PGQ)

The PGQ is a validated tool that captures self-reported pain and disability specifically for pregnant and post-partum women.[21,22] It has two subscales capturing activity limitations and symptoms. Items are scored on a 4-point descriptive scale, and item scores are summed and transformed to yield a score of 0 to 100, where 100 is the worst possible score. The PGQ has been shown to have acceptable responsiveness in women with PGP, LBP, or both.[14]

[H3]Numerical rating scale (NRS) for pain

The NRS asks patients to rate their pain severity on an 11-point scale where 0 indicates no pain and 10 indicates worst imaginable pain. Pain severity was measured using the mean of three 0 to 10 NRS for least, usual and current pain over the previous 2 weeks.[23]

[H2]Analysis

All analyses were performed using Stata version-14.[24] For a better comparison of the measures with differing scoring intervals, all scales were transformed to cover an interval ranging from 0–100. Change in scores was obtained by subtracting the 8 week follow-up scores from the baseline scores. Descriptive statistics (mean and standard deviations (SDs) or frequencies and percentages, as appropriate) were used to summarise the characteristics of participants who returned questionnaires at baseline and those who returned both baseline and follow-up questionnaires and to summarise the distribution of the change scores for the different outcomes according to whether the patients were classified as improved. Box plots were used to visualise the distribution of change score by the categories of GPE scale. The proportion of missing data at the item and scale level was computed. Since the percentage of missing (or not applicable) data for the questionnaire items was very minimal, the scores were recalculated as a percentage of the remaining items whenever an item was missing as recommended by the questionnaire developers; hence all the participants followed up had a score on all the instruments. Floor and ceiling effects were considered present if more than 15% of respondents achieved the lowest or highest possible score, respectively.[25]

The precision of the measures was evaluated using standard error of measurement (SEM),[26,27] calculated as $SD\sqrt{(1-r)}28$, where SD is the standard deviation of the baseline scores and r is the test-retest reliability coefficient obtained from a previous similar study.[22] The minimum detectable

change (MDC), defined as the lowest change, for an individual, that exceeds measurement error and noise at a 95% confidence level was calculated as $1.96*\sqrt{2}*SEM=2.77*SEM$.[28]

[H2]Responsiveness

Due to variations in method of assessing responsiveness, we adhere to the COnsensus based Standards for the selection of health Measurement Instruments (COSMIN)[10] recommendations in this paper. First, we determined the correlation between the change score and the rank of the self-rated GPE scale using Spearman's coefficient, where we hypothesised that there would be moderate-to-high correlations between change score and the GPE scale if they measure the same construct. Second, we used a receiver operating characteristic (ROC) curve analysis to plot the sensitivity versus 1-specificity for multiple cut-off points of the outcome measures against the external criterion. In the absence of a gold standard, the GPE scale, dichotomised into "improved" (combining categories 1 and 2) and not improved (combining categories 3-6) was used as the external criterion since this categorisation has been recommended as the optimal cut-off.[29] This measure, despite criticisms that it is a retrospective measure of change[30] and is influenced by current status,[31] has been shown to provide reliable assessments of health transition in people with musculoskeletal pain.[31] An area under the ROC curve (AUC) was used to assess the ability of the instruments to discriminate between patients who are considered to be improved and patients who are not considered to be improved according to the external criterion.[32]

[H2]MIC

The MIC was assessed using the anchor based approach to give an indication of the importance of the observed change by participating women. We used the ROC curve analysis where each instrument was considered as the diagnostic test, and the anchor as the gold standard, to distinguish patients with important improvement from patients with no important change.[33] We calculated sensitivity and specificity for each possible cut-off value of the change score and obtained an ROC curve. The MIC was defined as the optimal ROC cut-off point, that is, the change in score that is associated with the smallest amount of misclassification.

[H2]ROLE OF THE FUNDING SOURCE: The EASE Back study was funded by the National Institute for Health Research (NIHR) Health Technology Assessment Programme (HTA) (grant ref. no. 10/69/05) and supported by an NIHR Research Professorship to Professor Foster (NIHR-RP-011-015), who is an NIHR Senior Investigator. The funders played no role in the conduct, analysis, or reporting of this study, and the views expressed are those of the authors and not necessarily those of the National Health Service, the NIHR, or the UK Department of Health.

[H1]Results

[H2]Baseline characteristics

A total of 124 pregnant women (mean (SD) age 28 (5.3) years) were randomised and provided baseline data and 90 provided follow-up data at 8 weeks for all 3 measures. Characteristics of these participants have previously been reported.[15] As described in the report, a comparison of key baseline characteristics (age and index of multiple deprivations) captured from the anonymised data on women who were ineligible or declined participation with the EASE Back trial participants showed that there was reasonable overall comparability in these key baseline characteristics, indicating that our study sample was generally representative. Only 14% of participants had given birth by the time they returned the 8-weeks follow-up questionnaire. Table 1 shows the characteristics of the participants who completed questionnaires at baseline and those who provided data at both baseline and at 8-weeks follow-up. The subset of patients at 8-weeks follow-up was similar to the baseline sample in age, gestation week, pain location, pain duration, pain severity, and disability scores, but had a higher proportion of patients in full/part-time jobs and a slightly higher proportion of patients who were married.

[H2]Missing data, floor and ceiling effect

There were no missing data for pain severity, but minimal amounts of missing data for the ODI and PGQ. The proportion of the questionnaire items with missing data for all participants at baseline was 2.8% for ODI[1] and 6.2% for the PGQ. At follow-up the proportion was 7.8% and 8.7% for the ODI and PGQ, respectively. No floor or ceiling effects were shown in any of the measures– the total score for ODI ranged from 10% (scored by n=1) to 66% (scored by n=1) while the total PGQ score ranged from 13.6 (scored by n=1) to 85.3 (scored by n=1).

[1] The percentage is for the item missing data, for example, 35 items were missing out of a total of 1240 ODI scale items for the 124 participants. Similarly, there were 191 missing items (6.2%) out of possible 3100 items (25 x 124) for PGQ.

[H2]Change score from baseline

Table 2 shows the mean (SD) change score for improved and unimproved patients for all the three outcome measures.

Figure 1 shows the box plots of change scores for categories of GPE. Patients who reported complete recovery or much improved showed the highest change in all the measures while those who reported worsening showed deterioration on all the measures. Hence, change scores on all the measures corresponded to the magnitude of change as rated by patients on the GPE.

[H2]Responsiveness

The Spearman's correlation coefficient between the change score and the GPE scale for each measure was moderate to high ranging from 0.59 for PGQ activity subscale to 0.77 for pain severity (Tab. 3). All the measures showed good responsiveness (Fig. 2), with AUCs ranging between 0.77 and 0.90, indicating that all these measures are responsive over time in this patient population. Pain severity and the PGQ-symptoms subscale showed higher AUC compared to other measures**.** Exploratory subgroup analyses of the correlations by trial arms (Supplementary Tab. 1) and gestational weeks (Supplementary Tab. 2) showed no major differences between the study arms (especially between the standard care and true acupuncture) or gestational weeks, showing that randomisation and gestational week had no major effect on the measurement properties.

[H2]SEM, MDC and MIC estimates

Table 4 presents the SEM, MDC_{95} values, and the MIC identified by the optimal cut-off change in the ROC analysis together with the sensitivity and specificity at the cut-off for each outcome measure. The MIC ranged from 3.1 for ODI to 13.3 for pain severity (0–100 scale). When we relate the calculated MICs to the imprecision of the outcome measures assessed by the SEM (as we would want the MIC to

be at least higher than the SEM), we observe that the obtained MICs were all larger than the SEM except for the ODI.

[H1]Discussion

[H2]Summary of findings

This study provided a head-to-head comparison of the responsiveness of three PROMs, a 0-10 NRS for pain severity, the ODI and PGQ, and estimated the MIC for each. We found that all the three outcome measures had acceptable responsiveness based on a moderate-to-high correlation with the anchor and an AUC of more than 0.7, with NRS having the highest responsiveness, followed by the PGQ-symptoms subscale. Missing data in the three outcome measures was very minimal. No floor or ceiling effects were observed, indicating that the PROMs may be able to distinguish patients with the lowest or highest possible score.

Our study also estimated the MICs for the ODI, PGQ-total score, PGQ-activity subscale, PGQ-symptom subscale and 0-10 NRS pain severity as 3, 11, 9, 13 and 1.3 points, respectively. This indicates that when used at individual level, a change score smaller than these MIC values should be regarded as insignificant and only a change above these values can be considered as ‘‘clinically relevant’’ change. However, it’s worth noting that the obtained MIC for the ODI was almost the same as the SEM value indicating that for this measure there might be a risk that an obtained MIC is an expression of the imprecision of the instrument, and not a true change.

[H2]Comparison with previous studies

Previous studies have compared the responsiveness of PROMs measuring activity limitations in people with non-specific LBP,[20,34] however, to the best of our knowledge, only one previous study[14] has evaluated the responsiveness and MICs of these questionnaires in pregnant women who report PGP. Key differences between that previous study[14] and our study include: (1) study design – the previous study was a prospective cohort of women recruited from maternity centres, compared to our randomised trial comparing effectiveness of treatments in women seeking treatment for their pregnancy-related LBP and PGP; (2) study population – our study recruited women seeking treatment for pregnancy-related LBP and PGP who were between 13 to 31 weeks pregnant whereas Stuge et al’s[14] study recruited a consecutive sample of women from maternity centres in the last trimester of

pregnancy whose status was changing from pregnant to post-partum, irrespective of whether they had pain or were seeking treatment for it; and (3) timing of outcome assessment – our follow-up time-point was deliberately chosen to capture change in pain and function whilst women were still pregnant, whereas Stuge et al's[14] follow-up time-point was post-partum. Hence our study addresses a different question than the previous study,[14] as we focused on the responsiveness and MIC of the three outcome measures in women seeking treatment for pregnancy-related LBP and PGP, we included a different population of women and used a time-point during pregnancy in our evaluation of responsiveness and MIC. The latter point is important given the relevance to future clinical trials of treatment effectiveness of interventions for pregnancy-related LBP and PGP.

Estimates of MIC have previously been reported for the ODI in non-specific LBP patients[29,35] and for the ODI and PGQ in pregnant women transitioning to the post-partum period.[14] We found the MIC for the ODI to be 3.1 points. This value is lower than the MIC values previously reported in the non-specific LBP literature, where MICs of at least 10 points have been recommended.[29] This suggests that the ODI might not be a sensitive measure of activity limitations among pregnant women. Our MIC values for the PGQ and its subscales were also lower than the ones found for women proceeding from being pregnant to post-partum[14]. These differences in the MIC values may be explained by the fact that MIC cannot be considered a fixed property of an outcome measure[36] as it depends on the setting of the study in which the measure is used. Hence for our study population, an important change for pregnant women is not necessarily the same as an important change for the women as they transition from being pregnant to after giving birth. For pregnant women, activity limitation is often accepted as a 'normal' part of pregnancy, therefore, even a small improvement can be clinically meaningful to them, this may explain the lower MIC values obtained in our study.

The MIC value for pain severity measured by the NRS in non-specific LBP has also been previously reported in the literature to vary widely from 0.7 to 4 points.[29,37,38] However, there are no studies that have evaluated the MIC for the NRS for pain severity specifically in pregnant women, so we cannot make a direct comparison. Our MIC value for NRS pain severity of 1.3 points falls within the range for the general non-specific LBP population suggested by other authors.[29,37]

[H2]Implications

Despite having many outcome measures with which to assess non-specific LBP, currently there is no measure specifically for pregnancy-related LBP and PGP so comparing three measures and recommending the most appropriate one provides guidance for the future assessment of these pain presentations. The MIC is not a fixed attribute and usually fluctuates based on what is interpreted as important to the patient,[26] therefore, it is important to derive these values from the population of interest. Condition specific outcomes have been shown to have several advantages as they are intended to have very relevant content for the specific problem in that population and are therefore more likely to detect important changes that occur over time.[39] Our findings provide evidence regarding which outcome measures to recommend for assessing pain and function in women with pregnancy-related LBP and PGP. All the three measures showed acceptable responsiveness in our study. However, the MIC for the ODI was of similar magnitude to the measurement error thus questioning its use.

[H2]Strengths and limitations

A limitation of this study, which also applies to nearly all the responsiveness and MIC studies,[40,41] is the lack of a universally accepted methodology to determine responsiveness and MIC, which can result in a wide range of reported MIC values for a single outcome measure depending on the methodology chosen. However, we have adhered to the COSMIN recommendations which attempt to unify the methods for evaluating responsiveness and MIC. A further limitation of our study is the use of the GPE scale as the sole external criterion as there is no universally acceptable definition of "minimally important" using this anchor; hence our MICs would differ if a different anchor was chosen. However, it has been argued that for the GPE scale, the MIC is most appropriately defined in terms of at least "much improved" instead of including "somewhat improved,"[29] and this is the approach we took. Reassuringly, our results showed that change scores on all three outcome measures corresponded to the magnitude of change as perceived by patients. Further limitations of our study involve the inclusion of women reporting pain in areas consistent with a presentation of LBP, with or without PGP (as this was the population stipulated by the research funder), therefore, it is possible that the findings for the PGQ may be different than if we had included a sample of women who had PGP only;

however, there is no reliable method to distinguish clearly between these patient groups and in clinical practice as in this study, women present with both.

[H2]Future research

Since the number of patients who reported being unchanged was small, we relied on the reliability parameters from a previous study to calculate the SEM. A future test-retest study with a similar population should be carried out in order to confirm our findings. A future study with larger sample size could use other anchor based methods such as the mean change method based on patients who identify themselves to be ‘‘somewhat improved’’ on the patient-reported GPE scale.

[H1]Conclusion

Our analysis, comparing the ODI, PGQ and pain NRS, showed that all three outcome measures were responsive to change in women with pregnancy-related LBP and PGP and hence are suitable for measuring the effectiveness of treatments; of the three tested, the NRS for pain severity was the most responsive. The MIC of the ODI was comparable to its imprecision, hence its utility may be questionable. We identified the PGQ at 8 weeks post-randomisation as an appropriate outcome measure for pregnancy-related LBP and PGP. The NRS for pain severity is an efficient, shorter alternative.

AUTHOR CONTRIBUTIONS AND ACKNOWLEDGMENTS
The authors contributed to the manuscript as follows: conception and design: RO, AB, ML, MG, NEF; analysis and interpretation of the data: RO, AB, ML, MG, NEF; first drafting of the article: RO; revisions and final approval of the article: All.

Concept/idea/research design: R. Ogollah, A. Bishop, M. Lewis, M. Grotle, N.E. Foster
Writing: R. Ogollah, A. Bishop, M. Lewis, M. Grotle, N.E. Foster
Data collection: A. Bishop, N.E. Foster
Data analysis: R. Ogollah, A. Bishop, M. Grotle, N.E. Foster
Project management: A. Bishop, M. Lewis
Fund procurement: A. Bishop, N.E. Foster
Consultation (including review of manuscript before submitting): M. Lewis

The authors acknowledge funding for this study from the National Institute for Health Research (NIHR) Health Technology Assessment Programme (HTA), support from managers and clinicians in the University Hospital of the North Midlands (UHNM), the Staffordshire and Stoke-on-Trent Partnership NHS Trust; the CaRE for Women’s Health Network in the West Midlands of England, the NIHR Clinical Research Network (CRN), and participating clinicians and pregnant women. The authors also thank the research midwifery team at UHNM and the research nurse team within the CRN

for their help in identifying, recruiting, and conducting follow-up with women; the physical therapists who delivered the interventions; and the physical therapy administrators who organized EASE Back study appointments.

ETHICS APPROVAL

Ethics approval for the EASE Back pilot trial was granted by the National Research Ethics Service (NRES) Committee, West Midlands, Staffordshire (ref. no. 13/WM/00).

FUNDING

The EASE Back study was funded by the National Institute for Health Research (NIHR) Health Technology Assessment Programme (HTA) (grant ref. no. 10/69/05) and supported by an NIHR Research Professorship to Professor Foster (NIHR-RP-011-015), who is an NIHR Senior Investigator.

CLINICAL TRIAL REGISTRATION

This study used data from a pilot randomized clinical trial registered in the ISRCTN registry (ISRCTN49955124).

DISCLOSURES AND PRESENTATIONS

The authors completed the ICJME Form for Disclosure of Potential Conflicts of Interest and reported no conflicts of interest. The only publications arising from this work are in form of abstracts presented as posters at the International Back and Neck Pain Forum, Buxton, Derbyshire, UK, in June 2016. The full details of the EASE Back feasibility and pilot study have been published previously in an NIHR HTA monograph: Foster NE, Bishop A, Bartlam B, et al. Evaluating acupuncture and standard carE for pregnant women with back pain (EASE Back): a feasibility study and pilot randomized trial. *Health Technol Assess*. 2016;20:1-236.

References

1. Vermani E, Mittal R, Weeks A. Pelvic girdle pain and low back pain in pregnancy: a review. *Pain Practice.* 2010;10:60-71.

2. Liddle SD, Pennick V. Interventions for preventing and treating low-back and pelvic pain during pregnancy. *Cochrane Database Syst Rev.* 2015;(9):CD001139.

3. Mogren I. Perceived health, sick leave, psychosocial situation, and sexual life in women with low-back pain and pelvic pain during pregnancy. *Acta Obstetricia et Gynecologica Scandinavica.* 2006;85:647-656.

4. Vleeming A, Albert H, Östgaard H, Sturesson B, Stuge B. European guidelines for the diagnosis and treatment of pelvic girdle pain. *Eur Spine J.* 2008;17:794-819.

5. Wu WH, Meijer OG, Uegaki K, et al. Pregnancy-related pelvic girdle pain (PPP), I: terminology, clinical presentation, and prevalence. *Eur Spine J.* 2004;13:575-589.

6. Pierce H, Homer CS, Dahlen HG, King J. Pregnancy-related lumbopelvic pain: listening to Australian women. *Nurs Res Pract.* 2012;2012:387428.

7. Chiarotto A, Deyo RA, Terwee CB, et al. Core outcome domains for clinical trials in non-specific low back pain. *Eur Spine J.* 2015;24:1127-1142.

8. Chiarotto A, Boers M, Deyo RA, et al. Core outcome measurement instruments for clinical trials in nonspecific low back pain. *Pain.* 2018;159:481-495.

9. Boissonnault JS. A review of self-report functional outcome measures in selected obstetric physical therapy interventions. *J Womens Health Phys Therap.* 2009;33:7-12.

10. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol.* 2010;63:737-745.

11. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials.* 1989;10:407-415.

12. de Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. *Health Qual Life Outcomes.* 2006;4:54.

13. de Vet HCW, Terwee CB, Mokkink LB, Knol DL. *Measurement in Medicine: A Practical Guide*: Cambridge University Press; 2011.

14. Stuge B, Jenssen HK, Grotle M. The pelvic girdle questionnaire: responsiveness and minimal important change in women with pregnancy-related pelvic girdle pain, low back pain, or both. *Phys Ther.* 2017;97:1103-1113.

15. Foster NE, Bishop A, Bartlam B, et al. Evaluating Acupuncture and Standard carE for pregnant women with Back pain (EASE Back): a feasibility study and pilot randomised trial. *Health Technol Assess.* 2016;20:1-236.

16. Bishop A, Ogollah R, Bartlam B, et al. Evaluating acupuncture and standard care for pregnant women with back pain: the EASE Back pilot randomised controlled trial (ISRCTN49955124). *Pilot Feasibility Stud.* 2016;2:72.

17. Arain M, Campbell MJ, Cooper CL, Lancaster GA. What is a pilot or feasibility study? A review of current practice and editorial policy. *BMC Med Res Methodol.* 2010;10:67.

18. Fairbank JC, Couper J, Davies JB, O'Brien JP. The Oswestry low back pain disability questionnaire. *Physiotherapy.* 1980;66:271-273.

19. Fairbank JC, Pynsent PB. The Oswestry Disability Index. *Spine (Phila Pa 1976).* 2000;25:2940-2952; discussion 2952.

20. Davidson M, Keating JL. A Comparison of five low back disability questionnaires: reliability and responsiveness. *Phys Ther.* 2002;82:8-24.

21. Stuge B, Garratt A, Krogstad Jenssen H, Grotle M. The pelvic girdle questionnaire: a condition-specific instrument for assessing activity limitations and symptoms in people with pelvic girdle pain. *Phys Ther.* 2011;91:1096-1108.

22. Grotle M, Garratt AM, Krogstad Jenssen H, Stuge B. Reliability and construct validity of self-report questionnaires for patients with pelvic girdle pain. *Phys Ther.* 2012;92:111-123.

23. Dunn KM, Jordan KP, Croft PR. Recall of medication use, self-care activities and pain intensity: a comparison of daily diaries and self-report questionnaires among low back pain patients. *Prim Health Care Res Dev.* 2010;11:93-102.

24.*Stata Statistical Software: Release 14* [computer program]: College Station, TX: StataCorp LP; 2015.

25. Terwee CB, Bot SD, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol.* 2007;60:34-42.

26. Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. *J Clin Epidemiol.* 1999;52:861-873.

27. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. *J Clin Epidemiol.* 2003;56:395-407.

28.de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. *J Clin Epidemiol.* 2006;59:1033-1039.

29. Ostelo RW, de Vet HC. Clinically important outcomes in low back pain. *Best Pract Res Clin Rheumatol.* 2005;19:593-607.

30. Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. *J Clin Epidemiol.* 1997;50:869-879.

31. Kamper SJ, Ostelo RW, Knol DL, Maher CG, de Vet HC, Hancock MJ. Global perceived effect scales provided reliable assessments of health transition in people with musculoskeletal disorders, but ratings are strongly influenced by current status. *J Clin Epidemiol.* 2010;63:760-766 e761.

32. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures statistics and strategies for evaluation. *Control Clin Trials.* 1991;12:S142-S158.

33. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. *J Chronic Dis.* 1986;39:897-906.

34.van der Roer N, Ostelo RW, Bekkering GE, van Tulder MW, de Vet HC. Minimal clinically important change for pain intensity, functional status, and general health status in patients with nonspecific low back pain. *Spine (Phila Pa 1976).* 2006;31:578-582.

35. Ostelo RW, Deyo RA, Stratford P, et al. Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change. *Spine (Phila Pa 1976).* 2008;33:90-94.

36. Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PM. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. *Qual Life Res.* 2003;12:349-362.

37. Maughan EF, Lewis JS. Outcome measures in chronic low back pain. *Eur Spine J.* 2010;19:1484-1494.

38. Salaffi F, Stancati A, Silvestri CA, Ciapetti A, Grassi W. Minimal clinically important changes in chronic musculoskeletal pain intensity measured on a numerical rating scale. *Eur J Pain.* 2004;8:283-291.

39. Patrick DL, Deyo RA. Generic and disease-specific measures in assessing health status and quality of life. *Med Care.* 1989;27(3 Suppl):S217-S232.

40.de Vet HC, Ostelo RW, Terwee CB, et al. Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. *Qual Life Res.* 2007;16:131-142.

41. Terwee CB, Roorda LD, Dekker J, et al. Mind the MIC: large variation among populations and methods. *J Clin Epidemiol.* 2010;63:524-534.

Table 1: Baseline Characteristics of Participants Who Returned Questionnaires at Baseline and Those Who Provided Data at Both Baseline and at 8-Weeks Follow-up[a]

Baseline Characteristics	**Participants Randomized and Providing Baseline Data (n = 124)**	**Participants Returning Complete 8-Weeks Follow-up Questionnaire on All 3 Outcome Measures (n = 90)**
Age (years), mean (SD)	28.3 (5.3)	29.2 (5.3)
Highest qualification: degree/postgraduate, n (%)	45 (36.3)	40 (44.4)
Gestation weeks at inclusion: 24+, n (%)	52 (41.9)	37 (41.1)
Married, n (%)	54 (43.6)	46 (51.1)
Working (full-time or part-time), n (%)	87 (70.2)	71 (78.9)
Pre-pregnancy BMI, n (%)		
Normal/ underweight	47 (39.8)	37 (44.1)
Overweight	36 (30.5)	27 (32.1)
Obese/ morbidly obese	35 (29.7)	20 (23.8)
Pain location (manikin)[b], n (%)		
LBP only	23 (18.6)	17 (18.9)
LBP with anterior PGP	18 (14.5)	13 (14.4)
LBP with anterior PGP and pain elsewhere	36 (29.0)	22 (24.4)
LBP and pain elsewhere	47 (37.9)	38 (42.2)
Duration of episode: > 6 weeks, n (%)	68 (54.8)	46 (51.1)
Pain severity (mean of 3 0 - 10 NRS), mean (SD)	4.6 (1.7)	4.3 (1.6)
Oswestry Disability Index (0 - 100) [c], mean (SD)	34.4 (12.9)	32.3 (11.7)
Pelvic Girdle Questionnaire (0 - 100) [d], mean (SD)		
Total	53.3 (17.0)	51.6 (17.1)

Activity subscale	51.3 (17.9)	49.9 (18.2)
Symptom subscale	60.2 (18.4)	57.8 (17.9)

[a]LBP = low back pain; BMI = body mass index; PGP = pelvic girdle pain; SD = standard deviation
[b]LBP was defined as self-reported pain in the lumbar spine area (between the 12th rib and the gluteal fold); PGP was defined by pain between the posterior iliac crest and the gluteal fold, particularly in the vicinity of the sacroiliac joints; Pain elsewhere was pain not in the areas described by the definitions of LBP or PGP e.g. pain in the thoracic spine, pain in the shoulder, etc.
[c]Oswestry Disability Index- has 10 sections with scores ranging from 0 to 5 in each section, item scores are summed and transformed to yield a score of 0 to 100 where 100 is the worst possible score.
[d]Pelvic Girdle Questionnaire- items are scored on a 4-point scale and item scores are summed and transformed to yield a score of 0 to 100 where 100 is the worst possible score.

Table 2: Mean Change Scores (SD) for the 3 Outcome Measures According to Whether They Were Classified as Improved or Not Based on Their Answer on the GPE Scale.[a]

Key Outcomes	**Mean (SD) Change Score From Baseline**	
	Improved (n = 44)[b]	**Not Improved (n = 46)**
ODI score (0-100)	14.9 (12.2)	-3.1 (15.5)
PGQ (0-100)		
Total	21.8 (16.2)	-0.8 (21.6)
Activity subscale	20.1 (16.8)	-1.6 (22.9)
Symptom subscale	28.3 (19.2)	2.2 (21.0)
Pain severity (mean of 3 NRS)[b]	29.9 (15.6)	-2.8 (20.5)

[a]ODI = Oswestry Disability Index; PGQ = Pelvic Girdle Questionnaire; SD = standard deviation
[b]At 8 weeks follow-up, 6 participants (7%) reported complete recovery, 38 (42%) much improved, 17 (19%) somewhat improved, 7 (8%) the same, 11 (12%) somewhat worse and 11 (12%) much worse. Hence, 44 (49%) were classified as improved and 46 (51%) as not improved.

Table 3: AUC (With 95% CI), The Spearman Correlation Coefficient Between the Change Score and the Rank of the Self-Rated GPE Scale for Each Measure (N=90) and the Median Change Score Amongst Patients Who Rated Themselves as "Somewhat Improved"[a]

Key Outcomes	**Correlation with GPE Scale[b]**	**AUC (95% CI)**	**Median (IQR) Change Score for the "Somewhat Improved" GPE Category (n=17)**
ODI score (0-100)	-0.67	0.83 (0.74, 0.92)	8.0 (-6.2, 17.1)
PGQ (0-100)			
Total	-0.62	0.79 (0.70, 0.89)	11.0 (3.9, 29.2)
Activity subscale	-0.59	0.77 (0.67, 0.87)	8.8 (-1.5, 33.3)
Symptom subscale	-0.66	0.83 (0.74, 0.92)	13.3 (0.0, 20.0)
Pain severity (mean of 3 NRS)[b]	-0.77	0.90 (0.84, 0.97)	6.7 (0.0, 20.0)

[a]ODI = Oswestry Disability Index; PGQ = Pelvic Girdle Questionnaire; GPE = Global Perceived Effect; IQR = interquartile range; AUC = area under the curve; CI = confidence interval
[b]The negative sign for the correlation coefficient indicates that an increase in change score (improvement) for each of the measures is associated with lower rating (better improvement) on the GPE scale.

Table 4: Measurement Error, MDC, and MIC Values for Each of the Three Outcome Measures

Key Outcome Measures	ROC		MIC	SEM	MDC[b]
	AUC (95% CI)	Optimal Cut-Point	Sensitivity and Specificity at Cut-Point		
ODI score (0-100)	0.83 (0.74, 0.92)	3.1	0.89; 0.74	3.1	8.68
PGQ (0-100)					
Total	0.80 (0.70, 0.89)	11.0	0.77; 0.74	4.5	12.46
Activity subscale	0.77 (0.67, 0.87)	9.4	0.77; 0.72	4.7	13.12
Symptom subscale	0.83 (0.74, 0.92)	13.3	0.84; 0.72	5.5	15.29
Pain severity (mean of 3 NRS)[c]	0.90 (0.84, 0.97)	13.3	0.87; 0.85	5.2	20.44

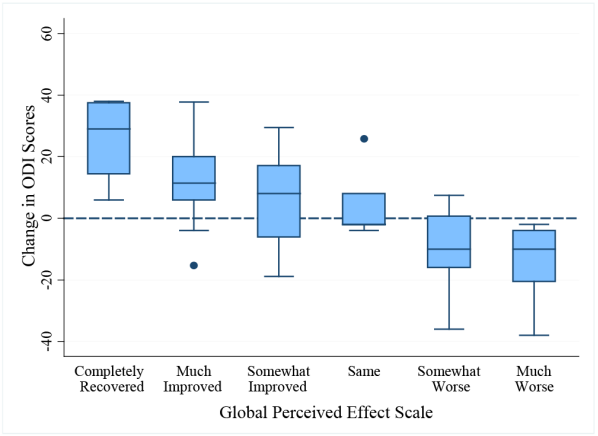
[a]AUC = area under the curve; CI = confidence interval; MIC = minimal important change; MDC = minimum detectable change ODI = Oswestry Disability Index; PGQ = Pelvic Girdle Questionnaire; ROC = receiver operating characteristic; SEM = standard error of measurement.

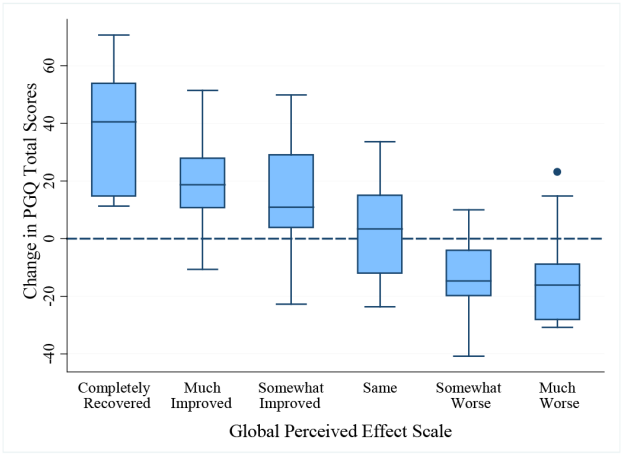
[b]The MDC_{95} represents the smallest change score that could be said to represent a real change beyond measurement error with 95% confidence in one individual. For example, the MDC_{95} value for ODI indicates that in 95% of the cases, patients will have experienced real change (beyond measurement error) if their score has changed by at least 8.68 points on the ODI.

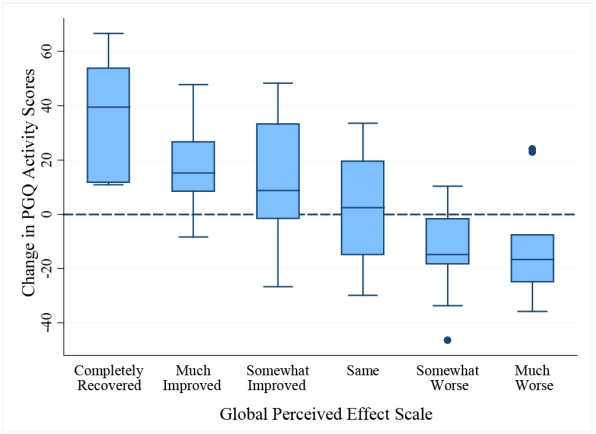
[c]The NRS has been transformed to 0-100 for comparability with other outcome measures. For SEM computation, a reliability of 0.94 was used for ODI, 0.9 for pain severity, 0.93 for PGQ total, 0.93 for PGQ-activity and 0.91 for PGQ-symptoms subscales obtained from a previous validation study.

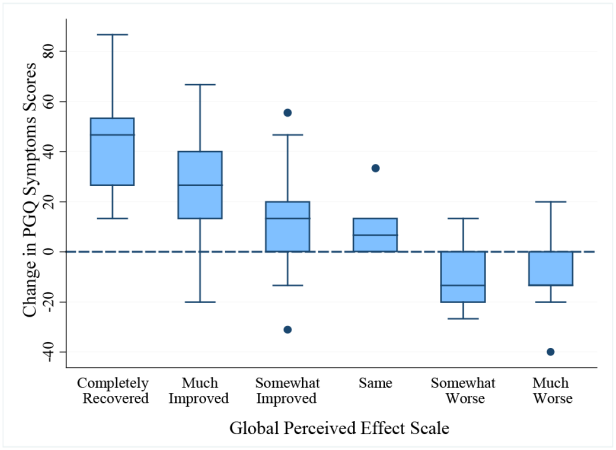
Figure Legend

Figure 1: Box plots of the distribution of change scores (baseline minus follow-up) for the different outcome measures in relation to categories of GPE scale. ODI = Oswestry Disability Index; PGQ = Pelvic Girdle Questionnaire.









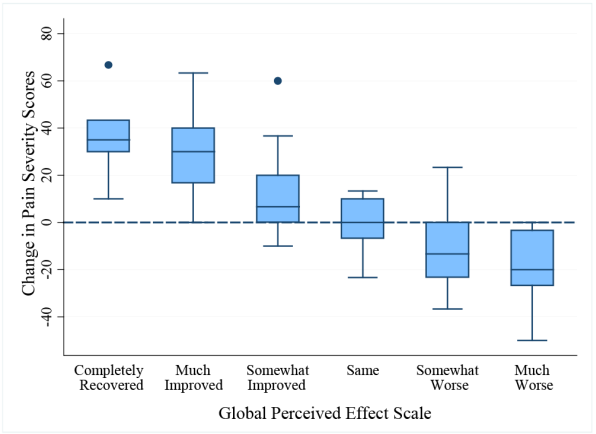


Figure 2: The receiver operating characteristic (ROC) curves of key outcome measures against global perceived effect. AUC = area under the curve; ODI = Oswestry Disability Index; PGQ = Pelvic Girdle Questionnaire.